VI. An attempt to ascertain the Prevalence of Syphilis in a large Urban Population.

Notice of FRITZ LENZ: Über die Verbreitung der Lues, speziell in Berlin, und ihre Bedeutung als Faktor des Rassentodes. *Archiv für Rassen- und Gesellschafts-Biologie.* May and June, 1910. Leipzig. B. G. Teubner. pp. 306 *et seq.*

The underlying ideas of the memoir are to use (1) the relative statistics in two places, and (2) the death rates from certain causes in order to estimate the number of persons attacked by the diseases which end in death from those causes. These ideas are excellent, and ultimately many valuable results may be reached, but they are extremely difficult to apply without making assumptions so wide that the conclusions become too rough to afford definite information. The particular case dealt with by Lenz is that of syphilis, which results in many cases in general paralysis of the insane, locomotor ataxia, etc., and the statistical problem involved in his work may be set out as follows: Given that syphilis is notified in Copenhagen, and that the deaths from general paralysis in Copenhagen and Berlin are known, find the proportion of males in Berlin who have at one time or another had syphilis. To solve such a problem one requires to know the age incidence of the deaths from paralysis, the age incidence of the syphilis notifications, the total populations in age groups (and the births) in both cities for several years, and some information as to the average after lifetime of syphilitics. Lenz neglects these preliminaries and boldly takes a short cut which assumes that if syphilis were notified in Berlin the notifications would bear the same proportion to the deaths from paralysis as the notifications in Copenhagen bear to the deaths from paralysis there. The syphilitic population is found by multiplying this number of notifications by the expectation of life at age 15. We have cut down Lenz's problem and have merely tried to indicate his method; he adjusts some details on the way, but the errors in the method we have just indicated exist, we think, in his work, though at times they are obscured. The weakness is that the proportionate method will not hold because the populations vary and the age incidence in the two cities can hardly be the same, while the use of the expectation at age 15 is incorrect, because this would be the youngest age at attack, and if expectation is used at all it should be for the average age of attack. Besides this the expectation of life of a syphilitic is probably less than that of the population as a whole.

These criticisms appear to us to dispose of his applications, but although the problem is an actuarial one of great difficulty it is certainly worth examination, and even though we do not agree with all his work we feel that much credit is due to Lenz for calling attention to the possibility of solving the problem of the extent of syphilis in this manner.

W. P. E.

VII. On the General Theory of the Influence of Selection on Correlation and Variation.

BY KARL PEARSON, F.R.S.

(1) In 1901 a paper of mine was read before the Royal Society and shortly afterwards issued in the *Philosophical Transactions** dealing with this matter. Very shortly afterwards I found out that the formulae therein developed did not depend for their accuracy on the frequencies being Gaussian in character. All the main conclusions were deducible without this limitation,

* Vol. 200 A, pp. 1—66.

and proofs of the formulae concerned have been given by me for many years in college lectures without making this assumption. I have, however, stated the fact once or twice in print that the formulae are general, and it seems desirable to reproduce my proofs (from lecture notes) at the present time as some doubt seems to have been cast on the generality of the formulae.

(2) I begin with a preliminary proposition, which is fairly familiar. Let $x_1, x_2 \dots x_n$ be n variables, continuous or discrete but quantitatively measurable; $\bar{x}_1, \bar{x}_2 \dots \bar{x}_n$ their means, $\sigma_1, \sigma_2, \dots \sigma_n$ their standard deviations, r_{pq} the correlation coefficient of the pth and qth variables supposed found by the product moment formula. Then: What is the best linear function u of $x_1, x_2 \dots x_n$ such that an $(n+1)$th variable x_{n+1} will have a maximum correlation $r_{ux_{n+1}} = \rho$, say, with u?

Let
$$u = c_1 x_1 + c_2 x_2 + \dots + c_n x_n$$
$$= S(cx).$$

Clearly
$$\bar{u} = S(c\bar{x}),$$
$$N\rho\sigma_u\sigma_{n+1} = \Sigma(u-\bar{u})(x_{n+1}-\bar{x}_{n+1}) = S\Sigma c_p(x_p-\bar{x}_p)(x_{n+1}-\bar{x}_{n+1})$$
$$= S\Sigma(c_p\sigma_p\sigma_{n+1}r_{p,n+1})$$
$$\rho = \frac{1}{\sigma_u} S(c_p\sigma_p r_{p,n+1}) \quad \dots\dots\dots\dots(i).$$
$$N\sigma_u^2 = \Sigma(u-\bar{u})^2 = S\Sigma\{c_p^2(x_p-\bar{x}_p)^2\} + 2S\Sigma c_p c_q (x_p-\bar{x}_p)(x_q-\bar{x}_q)$$
$$= NS(c_p^2\sigma_p^2) + 2NS(c_p c_q r_{pq}\sigma_p\sigma_q),$$

or:
$$\sigma_u^2 = S(c_p^2\sigma_p^2) + 2S(c_p c_q \sigma_p \sigma_q r_{pq}).$$

But if ρ be a maximum, we must have:
$$\frac{d\rho}{dc_p} = 0,$$
for all values of p.

Or, we reach type-equations of form:
$$\frac{1}{\sigma_u}\sigma_p r_{p,n+1} - \frac{\rho\sigma_u}{\sigma_u^2}\frac{d\sigma_u}{dc_p} = 0,$$

i.e.
$$r_{p,n+1} = \frac{\rho}{\sigma_p\sigma_u}\{c_p\sigma_p^2 + S(c_q\sigma_p\sigma_q r_{pq})\}$$
$$= \frac{\sigma\rho}{\sigma_u}(c_1\sigma_1 r_{1p} + c_2\sigma_2 r_{2p} + \dots + c_p\sigma_p + \dots + c_n\sigma_n r_{np})$$
$$= b_1 r_{1p} + b_2 r_{2p} + \dots + b_p + \dots + b_n r_{np} \quad \dots\dots\dots\dots(ii).$$

Where:
$$b_q = \rho\frac{c_q}{\sigma_u}\sigma_q.$$

Now equations of type (ii) are easily solved by aid of the determinant
$$R = \begin{vmatrix} 1, & r_{1,2}, & r_{1,3}, & \dots r_{1,n+1} \\ r_{2,1}, & 1, & r_{2,3}, & \dots r_{2,n+1} \\ \dots & \dots & \dots & \dots \\ r_{n+1,1}, & r_{n+1,2}, & r_{n+1,3}, & \dots 1 \end{vmatrix}.$$

We have
$$b_q = -R_{q,n+1}/R_{n+1,n+1}$$
where $R_{q,n+1}$ is the minor of qth row and $n+1$th column.

Hence it follows that:
$$u = -\frac{\sigma_u}{\rho}\left\{\frac{R_{1,n+1}}{R_{n+1,n+1}}\frac{x_1}{\sigma_1} + \frac{R_{2,n+1}}{R_{n+1,n+1}}\frac{x_2}{\sigma_2} + \dots + \frac{R_{n,n+1}}{R_{n+1,n+1}}\frac{x_n}{\sigma_n}\right\} \quad \dots\dots\dots\dots(iii).$$

Substitute also in (i):

$$\rho=\frac{1}{\rho}\sum_1^n(b_p r_{p,n+1}),$$

or:
$$\rho^2=-\sum_1^n\left(\frac{R_{p,n+1}}{R_{n+1,n+1}}r_{p,n+1}\right),$$

but:
$$R=r_{1,n+1}R_{1,n+1}+r_{2,n+1}R_{2,n+1}+\ldots+R_{n+1,n+1}.$$

Hence:
$$\rho^2=1-\frac{R}{R_{n+1,n+1}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(iv)},$$

$$1-\rho^2=1+\sum_1^n\frac{(R_{p,n+1}r_{p,n+1})}{R_{n+1,n+1}}=\frac{R}{R_{n+1,n+1}}.$$

Hence we have the following results:

$$\rho=\pm\sqrt{1-\frac{R}{R_{n+1,n+1}}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(v)},$$

$$u=-\frac{\sigma_u}{\rho}\left\{\sum_1^n\frac{R_{p,n+1}}{R_{n+1,n+1}}\frac{x_p}{\sigma_p}\right),$$

and
$$\sigma_{n+1}\sqrt{1-\rho^2}=\sigma_{n+1}\sqrt{\frac{R}{R_{n+1,n+1}}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(vi)},$$

and is the reduced average variability of x_{n+1} for given values of $x_1, x_2, \ldots x_n$.

The probable value of x_{n+1} is given by

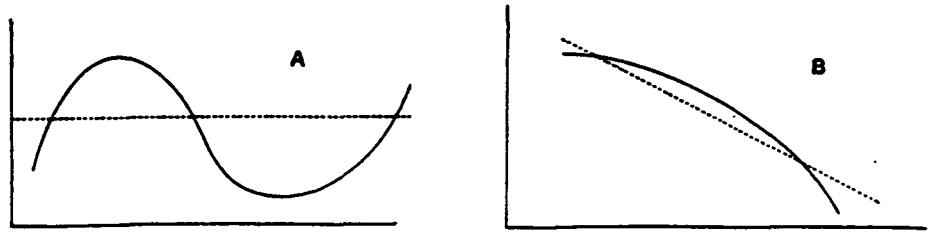
$$x_{n+1}-\bar{x}_{n+1}=\frac{\sigma_{n+1}}{\sigma_u}\rho(u-\bar{u})$$

$$=-\sum_1^n\frac{R_{p,n+1}}{R_{n+1,n+1}}\frac{\sigma_{n+1}}{\sigma_p}(x_p-\bar{x}_p) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(vii)},$$

i.e. the ordinary multiple regression formula. It is the "best value," i.e. the mean value of x_{n+1}, for given $x_1 \ldots x_n$, on the assumption that we correlate x_{n+1} with that *linear* function of the n variables, which gives the highest degree of relationship as measured by the correlation coefficient. The method is absolutely independent (i) of Gaussian theory, (ii) of the continuity or discreteness of the variables, but it does assume that linearity applies within the degree of useful approximation*.

Another point deserves re-emphasising here. Equation (iv) gives ρ^2, hence whether ρ be plus or minus, the errors of random sampling will always give a positive ρ^2. It follows therefore that even if ρ be zero, we should find on making a number of trials in each case a positive value of ρ^2; let the mean value of this be $\bar{\rho}^2$, then unless the actual value of ρ^2 is significant not as compared with zero, but with $\bar{\rho}^2$, no value ought to be laid on the actual value of ρ^2. The

* The general linearity ought to be tested in all such cases. Nothing can be learnt of association by assuming linearity in a case with a regression line (plane, etc.) like *A*, much in a case like *B*. To *A* we must apply multiple correlation-ratios, the theory of which is being developed at the present time and will shortly be published.



probable error of ρ is $\cdot 67449\,(1-\rho^2)/\sqrt{N} = \cdot 67449/\sqrt{N}$ if ρ be really zero; then if $\bar{\rho}$ be the mean value of ρ we should expect ρ to be

$$\bar{\rho} \pm \cdot 67449/\sqrt{N}$$

if ρ be truly zero. In other words we must consider the question of whether the observed ρ is significant compared with this.

I have found the value of $\overline{\delta\rho^2}$, i.e. the mean increment of ρ^2 due to errors of random sampling, but I postpone its consideration in the hope of still further reducing its determinantal expression in the general case.

Let us now apply these results to the general theory of selection. Suppose we have m variates $x_1, x_2, \ldots x_m$, with means $\bar{x}_1, \ldots \bar{x}_m$, standard deviations $\sigma_1, \sigma_2, \ldots \sigma_m$ and correlations given by R the determinant

$$\begin{vmatrix} 1 & r_{12} & r_{13} & \ldots & r_{1m} \\ r_{21} & 1 & r_{23} & \ldots & r_{2m} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{m1} & r_{m2} & r_{m3} & \ldots & 1 \end{vmatrix}$$

in the usual way.

Now we may suppose a selection to be made out of this variate complex of a subpopulation $x_1, x_2 \ldots x_n$, $n < m$, given by the means:

$$h_1, h_2, \ldots h_n,$$

the standard deviations

$$s_1, s_2, \ldots s_n$$

and the correlations:

$$\begin{vmatrix} 1 & \rho_{12} & \rho_{13} & \ldots & \rho_{1n} \\ \rho_{21} & 1 & \rho_{23} & \ldots & \rho_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \rho_{n1} & \rho_{n2} & \rho_{n3} & \ldots & 1 \end{vmatrix},$$

the selected population having values consistent with those of the unselected population.

We can then ask:

(i) How will this modify the mean and standard deviation of a non-selected variate x_p, $p > n < = m$?

(ii) How will this modify the correlation r_{pq} between two non-selected variates x_p and x_q, p and $q > n$ and $< = m$?

(iii) How will this modify the correlation r_{pt} of a non-selected and a selected variate $p > n$ and $< = m$, while $t < = n$? These are the fundamental problems of the influence of selection on variation and correlation.

(i) Let us take x_{n+1} as the non-selected organ and let the characters of one of the selected group be given by $x_t = \bar{h}_t + \xi_t$.

Then x_{n+1} will differ from its probable mean value by some quantity η_{n+1} and by (vii) we have

$$x_{n+1} = \bar{x}_{n+1} + \eta_{n+1} - \sum_1^n \left\{ \frac{R_{t,n+1}}{R_{n+1,n+1}} \frac{\sigma_{n+1}}{\sigma_t} (\bar{h}_t + \xi_t - \bar{x}_p) \right\}.$$

Or taking the mean value, $\bar{x}_{n+1}$, $S(\eta_{n+1}) = 0$ and $S(\xi_t) = 0$, and

$$\bar{x}_{n+1} = \bar{x}_{n+1} - \sum_1^n \left\{ \frac{R_{t,n+1}}{R_{n+1,n+1}} \frac{\sigma_{n+1}}{\sigma_t} (\bar{h}_t - \bar{x}_p) \right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(viii)}.$$

This establishes the first proposition* of my *Phil. Trans.* memoir, namely: that selection about the means with any variabilities gives the same mean value for a non-selected but correlated variate as if all the selected variates had been taken at their mean selected values.

We have clearly:

$$x_{n+1} - \bar{x}_{n+1} = \eta_{n+1} - \sum_1^n \left(\frac{R_{t,n+1}}{R_{n+1,n+1}} \frac{\sigma_{t+1}}{\sigma_t} \xi_t \right) \quad \text{.....(ix).}$$

Now if we are dealing with N manifolds of variates:

$S(\eta^2_{n+1})/N$ = a standard deviation indicated by $\bar{\sigma}_{n+1}$,

$S(\xi_t^2)/N$ = standard deviation of selected tth organ $= s_t^2$,

$S(\xi_t \xi_r)/N = s_t s_r \rho_{tr}$,

$S\{\eta_{n+1} \xi_t\} = 0$ because the tth variate is not selected in reference to the $(n+1)$th variate. Hence if we square (ix) and call Σ_{n+1} the resulting variability of x_{n+1} due to the selection of the n-variates, we have

$$\Sigma^2_{n+1} = \sigma^2_{n+1} \left\{ \frac{\bar{\sigma}^2_{n+1}}{\sigma^2_{n+1}} + \sum_1^n \left(\frac{R^2_{t,n+1}}{R^2_{n+1,n+1}} \frac{s_t^2}{\sigma_t^2} \right) + 2S \left(\frac{R_{t,n+1} R_{r,n+1}}{R^2_{n+1,n+1}} \frac{s_t s_r}{\sigma_t^2} \rho_{tr} \right) \right\}.$$

But as we have already seen η_{n+1} is not correlated with ξ_t. Hence we shall find the value of $\bar{\sigma}^2_{n+1}$ by putting all the s_t's zero, or by concentrating the selection at a single value of a manifold. It is therefore the value of Σ^2_{n+1} for an array of x_{n+1} for definite values of $x_1, x_2 \ldots x_n$, i.e. by (vi) $\bar{\sigma}_{n+1}$ equals $\sigma_{n+1} \sqrt{\dfrac{R_{(n+1)}}{R_{n+1,n+1}}}$, where $R_{(n+1)}$ is the determinant of $n+1$ rows and columns. Thus finally:

$$\Sigma^2_{n+1} = \sigma^2_{n+1} \left\{ \frac{R_{(n+1)}}{R_{n+1,n+1}} + \sum_1^n \left(\frac{R^2_{t,n+1}}{R^2_{n+1,n+1}} \frac{s_t^2}{\sigma_t^2} \right) + 2S \left(\frac{R_{t,n+1} R_{r,n+1}}{R^2_{n+1,n+1}} \frac{s_t s_r}{\sigma_t^2} \rho_{tr} \right) \right\} \quad \text{.....(x).}$$

This is in complete agreement with the value given as Equation (xlv) of my *Phil. Trans.* memoir†, and deduced there on the assumption of a Gaussian frequency distribution.

(ii) I now turn to the second of my problems the correlation between the $(n+1)$th and $(n+2)$th variables. In this work $R'_{n+2,n+2}$ denotes the determinant of n rows and columns bordered by the $(n+2)$th variate correlations, those of the $(n+1)$th being omitted. Clearly as in (ix)

$$x_{n+2} - \bar{x}_{n+2} = \eta_{n+2} - \sum_1^n \left(\frac{R'_{t,n+2}}{R'_{n+2,n+2}} \frac{\sigma_{n+2}}{\sigma_t} \xi_t \right) \quad \text{.....(xi).}$$

Multiply (ix) and (xi) sum and divide by the number of the manifolds, N; then if $\rho_{n+1,n+2}$ be the correlation after selection of the $(n+1)$th and $(n+2)$th variates, we have:

$$\Sigma_{n+1} \Sigma_{n+2} \rho_{n+1,n+2} = \frac{S(\eta_{n+1} \eta_{n+2})}{N} + \sigma_{n+1} \sigma_{n+2} \left\{ \sum_1^n \frac{R_{t,n+1} R'_{t,n+2}}{R_{n+1,n+1} R'_{n+2,n+2}} \frac{s_t^2}{\sigma_t^2} \right.$$

$$\left. + S \left(\frac{R_{t,n+1}}{R_{n+1,n+1}} \frac{R'_{r,n+2}}{R'_{n+2,n+2}} \frac{s_t s_r}{\sigma_t \sigma_r} \rho_{tr} \right) + S \left(\frac{R_{r,n+1}}{R_{n+1,n+1}} \frac{R'_{t,n+2}}{R'_{n+2,n+2}} \frac{s_t s_r}{\sigma_t \sigma_r} \rho_{tr} \right) \right\} \quad \text{.....(xii).}$$

As before $\dfrac{S(\eta_{n+1} \eta_{n+2})}{N}$ will be given by the mean partial product moment of the $(n+1)$th and $(n+2)$th variates for *constant* values of the n variates concentrated at their selected means. This can be found without appeal to the Gaussian frequency surface by extending the formula (vii) to $n+1$ variates.

* Vol. 200, A, p. 13.

† *Phil. Trans.* Vol. 200, A, p. 17.

Let Δ be the determinant of $(n+2)$ rows and columns, Δ_{pq} the minor corresponding to the pth column and qth row component. Then the regression equations for x_{n+1} and x_{n+2} on the remaining variates of the $(n+2)$ group are:

$$x_{n+1}-\bar{x}_{n+1}=-\frac{\Delta_{n+2,n+1}}{\Delta_{n+1,n+1}}\frac{\sigma_{n+1}}{\sigma_{n+2}}(x_{n+2}-\bar{x}_{n+2})-\sum_1^n\left\{\frac{\Delta_{t,n+1}}{\Delta_{n+1,n+1}}\frac{\sigma_{n+1}}{\sigma_t}(x_t-\bar{x}_t)\right\}$$

and

$$x_{n+2}-\bar{x}_{n+2}=-\frac{\Delta_{n+1,n+2}}{\Delta_{n+2,n+2}}\frac{\sigma_{n+2}}{\sigma_{n+1}}(x_{n+1}-\bar{x}_{n+1})-\sum_1^n\left\{\frac{\Delta_{t,n+2}}{\Delta_{n+2,n+2}}\frac{\sigma_{n+2}}{\sigma_t}(x_t-\bar{x}_t)\right\}.$$

Now, when we put $x_1 \ldots x_n$ constant, the coefficients of $x_{n+2}-\bar{x}_{n+2}$ and $x_{n+1}-\bar{x}_{n+1}$ and the partial regression coefficients of x_{n+1} on x_{n+2} and x_{n+2} on x_{n+1} for constant 1 to n variates, or the square root of their product is the partial correlation coefficient, i.e.

$$\rho_{n+1,n+2\,.\,1,2,3\ldots n}=\bar{\rho}_{n+1,n+2},$$

say for brevity; therefore

$$\bar{\rho}_{n+1,n+2}=\sqrt{\frac{\Delta^2_{n+1,n+2}}{\Delta_{n+1,n+1}\Delta_{n+2,n+2}}}=-\frac{\Delta_{n+1,n+2}}{\sqrt{\Delta_{n+1,n+1}\Delta_{n+2,n+2}}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\text{(xiv)},$$

a well-known and familiar form*.

Now let us look at the standard deviations of the arrays of the $(n+1)$th and $(n+2)$th variates for absolutely selected values of the n variates.

The variability of the array of the $(n+1)$ variate is given by (vi), i.e.

$$\bar{\sigma}_{n+1}=\sigma_{n+1}\sqrt{\frac{R}{R_{n+1,n+1}}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(xv)},$$

and of the $\overline{n+2}$th variate

$$\bar{\sigma}_{n+2}=\sigma_{n+2}\sqrt{\frac{R'}{R'_{n+2,n+2}}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(xvi)}.$$

But

$$R=\Delta_{n+2,n+2},$$

$$R'=\Delta_{n+1,n+1}$$

while clearly $R_{n+1,n+1}=R'_{n+2,n+2}=$ the second minor of Δ obtained by leaving out both $(n+1)$th and $(n+2)$th rows and columns. Hence we have:

$$R_{n+1,n+1}=R'_{n+2,n+2}=\Delta_{n+1,n+1,n+2,n+2},$$

and

$$\frac{\Sigma(\eta_{n+1}\eta_{n+2})}{N}=\bar{\sigma}_{n+1}\bar{\sigma}_{n+2}\bar{\rho}_{n+1,n+2}=-\sigma_{n+1}\sigma_{n+2}\frac{\Delta_{n+1,n+2}}{\Delta_{n+1,n+1,n+2,n+2}} \quad \ldots\ldots\text{(xvii)}.$$

Thus finally we have from (xii):

$$\Sigma_{n+1}\Sigma_{n+2}\rho_{n+1,n+2}=\sigma_{n+1}\sigma_{n+2}\left\{\frac{-\Delta_{n+1,n+2}}{\Delta_{n+1,n+1,n+2,n+2}}+\sum_1^n\left(\frac{R_{t,n+1}R'_{t,n+2}}{R_{n+1,n+1}R'_{n+2,n+2}}\frac{s_t^2}{\sigma_t^2}\right)\right.$$
$$\left.+S\left(\frac{R_{t,n+1}R'_{t',n+2}}{R_{n+1,n+1}R'_{n+2,n+2}}+\frac{R_{t',n+1}R'_{t,n+2}}{R_{n+1,n+1}R'_{n+2,n+2}}\right)\frac{s_ts_{t'}}{\sigma_t\sigma_{t'}}\rho_{tt'}\right\} \quad \ldots\ldots\ldots\text{(xviii)},$$

which is in complete agreement with the value found from the Gaussian hypothesis†.

(iii) Lastly we require the correlation $\rho_{t,n+1}$ between a selected and a non-selected variate, $t<=n$. Turning back to (ix) multiply by ξ_t, sum and divide by N, then:

$$s_t\Sigma_{n+1}\rho_{t,n+1}=\frac{S(\eta_{n+1}\xi_t)}{N}-\left(\frac{R_{t,n+1}}{R_{n+1,n+1}}\frac{\sigma_{n+1}}{\sigma_t}s_t^2\right)-S\left(\frac{R_{t',n+1}}{R_{n+1,n+1}}\frac{\sigma_{n+1}}{\sigma_{t'}}s_ts_{t'}\rho_{tt'}\right).$$

* Pearson, *Phil. Trans.* Vol. 200, A, p. 10, Equation (xxvii).

† See *Phil. Trans.* Vol. 200, A, p. 17, Equation (xlvi).

The first summation on the right is zero; hence

$$\Sigma_{n+1}\rho_{t,n+1} = -\sigma_{n+1}\left\{\frac{R_{t,n+1}}{R_{n+1,n+1}}\frac{s_t}{\sigma_t} + S\left(\frac{R_{t'n+1}}{R_{n+1,n+1}}\right)\frac{s_{t'}}{\sigma_{t'}}\rho_{tt'}\right\} \quad \text{.....(xix).}$$

This with a slight difference of notation is the result obtained on the Gaussian hypothesis*.

The above proofs justify the statement that the general selection formulae given by me are independent of any Gaussian assumption. They are really peculiar to the general idea of the manifold linear variate u which gives the maximum correlation coefficient of an $(n+1)$th variate with n other variates. They do not involve any idea of continuity or any hypothesis as to the nature of the selected means, standard deviations and correlations beyond the fundamental assumption that the selected population really exists inside the unselected population. There need be no hesitation therefore in applying these formulae to any cases whatever in which the correlation coefficients have valid application at all.

* *Phil. Trans.* Vol. 200, A, p. 17, Equation (xlvii). S in our present notation is a summation in (xix) of every value, but t, of t'. In the *Phil. Trans.* paper S_1 is a summation for *all* values of t': see p. 18.

VIII. On a Fallacious Proof of Sheppard's Correction.

The ordinary proofs of Sheppard's corrections for the moments are somewhat lengthy and depend entirely on the principle of high contact at the terminals. Mr G. U. Yule in his recent *Theory of Statistics*, p. 208, has given a proof in a few lines which is absolutely independent of this principle, and which from its very simplicity is likely, if not criticised, to be generally adopted. Unfortunately it is wholly fallacious. The error lies in the words "the correlation between X and δ is zero, for the mean value of δ is zero for every interval." What Mr Yule should have said is that the correlation between his Z and δ is zero, and he should have reached the conclusion

$$\sigma_1^2 = \sigma^2 + \tfrac{1}{12}c^2,$$

and not

$$\sigma_1^2 = \sigma^2 - \tfrac{1}{12}c^2,$$

for he is really working out the mean square for the histogram and not the true figure. He would thus have failed to obtain the correct value, which he does not appear to recognise arises solely from the fact that the 'trapezettes' *cannot* be treated as rectangles. In the case of curves of frequency without terminal contact, Sheppard's corrections are *not* the proper ones, and their general adoption without regard to their limitations is to be deprecated. Such adoption is directly encouraged by a fallacious proof of the above character. K. P.